

TUTORIAL ESTADÍSTICA DE VARIABLES CONJUNTAS

Es habitual encontrar en la industria y en mantenimiento con fenómenos físicos cuya explicación pasa por analizar el comportamiento de 2 variables íntimamente relacionadas. Por eso, el tutorial de esta semana se centrará en explicar qué funciones permiten determinar el grado de relación existente entre dichas variables. Por ejemplo, la relación entre las averías mensuales y su temperatura media de funcionamiento.

DISTRIBUCIONES DE FRECUENCIAS

Será preciso para analizar ese comportamiento conjunto disponer de datos experimentales, por lo que se definirá la *Distribución Conjunta de Frecuencias* a la tabla que representa los valores de las 2 variables indicadas, por ejemplo la mencionada con anterioridad:

Para el caso particular de relaciones de carácter cualitativo el nombre que se le da es el de *Tabla de Contingencias*.

AVERÍAS MENSUALES DE UNA MÁQUINA Y TEMPERATURA DE FUNCIONAMIENTO

La tabla disponible es la siguiente, obtenida estudiando el número total de averías producidas bajo diferentes condiciones de temperatura y calculando la frecuencia de fallo como:

$$fr(\text{suceso } A) = \frac{\text{Número de veces que se observa } A}{\text{Número total de datos}}$$

siendo un suceso, por ejemplo, 2 averías producidas con 40° en la máquina, calculándose su frecuencia como:

$$fr(2 \text{ averías con } 40^\circ) = \frac{\text{número de veces que se han producido averías a } 40^\circ}{\text{número total de averías}}$$

Habiéndose obtenido la tabla siguiente.

AVERÍAS	TEMPERATURA			SUMATORIA FILAS
	40°	50°	60°	
2	0,20	0,15	0,10	0,45
3	0,12	0,07	0,05	0,24
4	0,04	0,10	0,02	0,16
5	----	0,05	0,10	0,15
SUMATORIA COLUMNAS	0,36	0,37	0,27	1

REPRESENTACIÓN FUNCIONES MULTIVARIANTES

A cada uno de los valores que aparece en la tabla anterior se le llama:

$$fr(x_i, y_j)$$

Es decir, la frecuencia de la pareja de valores (x_i, y_j) , cumpliéndose:

$$\sum_i \sum_j fr(x_i, y_j) = 1$$

es decir, la suma de todas las casillas de una de esas tablas debe ser igual a la unidad.

DISTRIBUCIONES MARGINALES Y FRECUENCIAS RELATIVAS

Se define así a la distribución de datos que se obtiene estudiando una variable independientemente de la otra, o sea que en el caso anterior correspondería en fijar valores por filas o por columnas y realizar las sumatorias pertinentes obteniéndose:

$$fr(x_i) = \sum_j fr(x_i, y_j)$$

frecuencia relativa para cada valor de la variable X x_i , o alternativamente:

$$fr(y_j) = \sum_i fr(x_i, y_j)$$

frecuencia relativa para cada valor de la variable Y y_j

DISTRIBUCIONES CONDICIONADAS

Es aquella distribución univariante de Y para un valor $x=x_i$ siendo sus valores las sumas para cada x_i en lugar de los valores individuales. En ese caso las frecuencias relativas de la distribución se llamarían *frecuencias condicionadas* de valor:

$$fr(y_j|x_i) = \frac{fr(x_i, y_j)}{fr(x_i)}$$

Es decir, la frecuencia condicionada de un valor de Y y_j fijado el valor de X x_i sería el cociente entre la frecuencia absoluta del par por la relativa del valor x_i , verificando que:

$$\sum_j fr(y_j|x_i) = \frac{1}{fr(x_i)} \sum_j fr(x_i, y_j) = 1$$

Para la tabla anterior cuando x toma el valor de 5 averías tendríamos que.

$$fr(Temperatura\ 40\ |\ 5\ averías) = 0,05/0,15 = 0,33$$

y que:

$$fr(Temperatura\ 50\ |\ 5\ averías) = 10/0,15 = 0,67$$

VARIABLES CARACTERÍSTICAS

COVARIANZA

Se define como:

$$Cov(x, y) = \frac{\sum (x_i - \bar{x})(y_j - \bar{y})}{n}$$

o bien si se agrupan los valores:

$$Cov(x, y) = \sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y}) fr(x_i, y_j)$$

Este valor mide la relación lineal entre x e y.

CORRELACIÓN

Ante la existencia de posibles relaciones desfavorables entre las unidades de las variables analizadas resulta conveniente evaluar también la correlación definida como (adimensional):

$$r = \frac{Cov(x, y)}{s_x s_y}$$

donde s_x y s_y representan respectivamente las desviaciones típicas de las variables X e Y. Este valor, como se dijo adimensional, tomará valores igual a +/-1 si la relación es lineal exacta, y sino se situará en el intervalo $-1 < r < +1$.

REPRESENTACIÓN GRÁFICA

Es Conveniente, pese a analizar los valores anteriores, el representar gráficamente los valores pues no siempre se puede extrapolar la representación gráfica a partir de los indicadores anteriores.

En el caso de existir uan relación lineal se puede extrapolar su relación mediante una gráfica conocida como *recta de regresión*. En ese caso, dicha recta presentará la forma:

$$Y = aX + b$$

donde b tomará el valor siguiente:

$$b = \frac{Cov(x, y)}{s_x^2}$$

GENERALIZACIÓN

Todo lo anterior se puede generalizar para relaciones k-dimensionales mediante el uso de notación vectorial y matricial hablándose de:

- Vector de Medias $\bar{X} = \frac{1}{n} \begin{bmatrix} \bar{x} \\ \bar{y} \\ \dots \\ \bar{z} \end{bmatrix}$

- Matriz de Varianzas y Covarianzas $M = \begin{pmatrix} s_x^2 & s_{xy} & s_{xz} \\ s_{yx} & s_y^2 & s_{yz} \\ \dots & \dots & \dots \\ s_{zx} & s_{zy} & s_z^2 \end{pmatrix}$

donde $s_{xy} = s_{yx}$ la covarianza de las variables X e Y, y s_x, s_y las desviaciones típicas de X e Y respectivamente.

- Varianza Generalizada $M = \begin{pmatrix} s_x^2 & rs_x s_y \\ rs_x s_y & s_y^2 \end{pmatrix}$ para dos dimensiones siendo el determinante de la misma el área ocupada por los datos y una medida de la dispersión en sí misma.

EJEMPLO

3 variables que pueden estar muy relacionadas son:

- Tasa real del Crecimiento del Producto Nacional Bruto (PNB) en una década
- Tasa de Desempleo (década + 4 años)
- Tasa de Incremento de Precios al Consumo (década + 3 años)

Los datos se muestran en la siguiente tabla:

PNB DÉCADA (X)	DESEMPLEO DÉCADA +4 (Y)	DESEMPLEO DÉCADA + 3 (Z)
2,0	8,00	2,2
2,8	7,50	7,6
2,6	4,50	2,9
2,2	13,50	4,6
2,8	9,50	4,1
1,8	8,75	3,9
2,6	22,50	7,4
2,2	7,25	3,2
3,1	6,25	5,1
2,7	10,75	5,3
3,1	9,00	20,1
1,9	14,00	2,3
4,0	16,75	5,5
3,4	1,00	32,7
2,6	10,50	9,1
4,3	2,75	1,7
1,7	1,75	3,2
4,0	2,25	5,8
1,6	4,75	16,3
3,8	11,50	15,9
1,5	11,50	5,9
1,6	3,00	6,7
0,6	0,50	3,4
5,1	13,50	40,5

Para las 3 variables anteriores, ANALIZADAS INDEPENDIENTEMENTE, se tienen las siguientes medidas características:

	X	Y	Z
MEDIA	2,67	8,38	8,97
MEDIANA	2,60	8,00	5,30
D. TÍPICA	1,05	5,43	9,77

El análisis conjunto muestra las siguientes medidas:

	X	Y	Z
X	1,10 (s_x^2)	1,11 (s_{xy})	5,12 (s_{xz})
Y	1,11 (s_{xy})	29,50 (s_y^2)	2,12 (s_{yz})
Z	5,12 (s_{xz})	2,12 (s_{yz})	95,58 (s_z^2)

con los siguientes coeficientes de correlación:

- $r_{xy} = \frac{1,11}{1,05*5,43} = \mathbf{0,19}$
- $r_{xz} = \frac{5,12}{1,05*9,77} = \mathbf{0,50}$
- $r_{yz} = \frac{2,12}{5,43*9,77} = \mathbf{0,04}$

Luego la relación existe de manera más notable entre variables X (crecimiento) y Z (inflación), presentando las demás variables una menor relación.